
Lower Bounds for Multi-armed Bandit with Non-equivalent Multiple Plays

Aleksandr Vorobev
Yandex
Moscow, Russia
alvor88@yandex-team.ru

Gleb Gusev
Yandex
Moscow, Russia
gleb57@yandex-team.ru

Abstract

We study the stochastic multi-armed bandit problem with non-equivalent multiple plays where, at each step, an agent chooses not only a set of arms, but also their order, which influences reward distribution. In several problem formulations with different assumptions, we provide lower bounds for regret with standard asymptotics $O(\log t)$ but novel coefficients and provide optimal algorithms, thus proving that these bounds cannot be improved.

1 Introduction

Multi-armed bandit (MAB) is a common model to formulate problems of finding the tradeoff between exploration and exploitation. Its stochastic formulation with multiple plays was originally considered in [3]. In this formulation, at each step of a game, an agent chooses m arms from an arm set A and observes the reward for each of them, which is a random variable whose distribution is a property of the arm. The agent's goal is to minimize the expected cumulative *regret* over the first T steps, i.e., the difference between the expected cumulative reward of the observed arms for the optimal strategy, which relies on the complete information about the reward distributions of all the arms, and the chosen strategy, which relies on the past observations only. In the paper [3], theoretical analysis of the asymptotic behavior of the cumulative regret is provided.

An important limitation of [3] is that the rewards of the chosen arms are supposed to be independent of the order the agent put them into the set. In many applications, on the contrary, the same arm can exhibit different reward distributions at different positions. In particular, problems of web search ranking [16, 18], recommendations [13, 17], and contextual advertising [2, 14] are often formulated as MAB problems with documents, recommended items, and ads respectively as arms. Steps of the game correspond to the requests of users, the application (agent) chooses objects to show them in different slots (positions) of the web page, and the user's interaction with an object (which defines its reward) clearly depends on the slot of the page the object is placed in.

Some papers studied adversarial bandit settings with non-equivalent plays [11, 9, 6, 1]. Some other studies [16, 18, 15, 17] consider stochastic problem formulations and prove upper bounds for the regret of corresponding algorithms. All these algorithms follow a general scheme: they rank arms by some score which balances between exploration and exploitation, and choose the top arms for the slots in the order of the slots' importance. Thereby, these algorithms use the same exploration rate to choose arms for different positions. However, it follows from [3] that, in stochastic setting, even in the case of equivalent plays, an asymptotically optimal algorithm should explore only one arm at one step most part of time.

In this paper, we consider several settings of the general stochastic non-contextual MAB problem with non-equivalent multiple plays. These settings (see Section 2 for description) differ by additional restrictions on the parameter space of arms and the reward distributions of their lists. These assumptions were held in many above-mentioned works and handle a variety of application tasks.

In the chosen settings, we provide lower bounds for the asymptotic behavior of the cumulative regret in Section 3 and prove their tightness under additional reliable requirements by presenting an algorithm with the same regret asymptotic behavior. Importantly, the form of each lower bound gives an insight on the construction of optimal algorithms in some specific cases not covered by our algorithm.

2 Problem Formalization

Let us consider the following problem. There is a parameter space \mathbf{A} equipped with continuously distributed random vectors $F(\bar{a})$ with values in \mathbb{R}^d , densities $f(\cdot, \bar{a})$, and finite expectations $\mu(\bar{a})$ for each list $\bar{a} \in \mathbf{A}^m$ of values of a fixed length m . We require each component F_i of $F(\bar{a})$ to be integrable: $\int_{\mathbb{R}^d} |x_i| f(x, \bar{a}) dx < \infty$. The case, where all distributions $f(\cdot, \bar{a})$ are discrete, can be considered as well by substituting probability functions for densities $f(\cdot, \bar{a})$ and substituting summation for integration everywhere in the paper.

At the start, an agent is provided with the space \mathbf{A} and *arms* $1, 2, \dots, N$, where each arm j is provided with an unknown parameter $a_j \in \mathbf{A}$. We denote $A := (a_1, \dots, a_N)$. At each step t , the agent chooses a list of different arms $\pi_t = (\pi_t(1), \dots, \pi_t(m))$ ($\pi_t(j_1) \neq \pi_t(j_2)$ if $j_1 \neq j_2$) to fill a row of slots $S = \{1, \dots, m\}$ with them. We denote the set of all the lists of m different arms by Π . Next, the agent observes a realization F_t of $F(\bar{a}_{\pi_t})$ (F_t are independent over steps), where $\bar{a}_{\pi(t)} = (a_{\pi_t(1)}, \dots, a_{\pi_t(m)})$, and further utilizes it for choosing lists at future steps. Note that $f(\cdot, \bar{a})$ can be not invariant with respect to permutations, i.e., the order of the arms in the list is important. The agent's goal is to minimize the cumulative regret \mathbf{Reg}_T over the first T steps:

$$\mathbf{Reg}_T = T \max_{\pi \in \Pi} \mathbb{E}R(\bar{a}_\pi) - \mathbb{E} \sum_{t=1}^T R_t,$$

where $R(\bar{a}) = U(F(\bar{a}))$, $R_t = U(F_t)$, and $U : \mathbb{R}^d \mapsto \mathbb{R}$ is the function of reward depending on the observed values. Splitting the standard notion of an observed reward into the observed values F and the reward R allows to handle the case of observing only the list reward (see [6, 1]) as well as the cases when the agent observes a contribution of each individual arm to the list reward (see, e.g., Assumption 2) or other aspects of the interaction that can provide additional information on a_j , e.g., the time to the first click or the session duration in the case of web services. The described problem setting generalizes the one considered in [3] to non-equivalent plays and a more general form of relation between observed values F and the optimized reward R .

3 Lower Bounds for Regret

Before presenting each of our results, we introduce some notations and additional assumptions (on the space $(\mathbf{A}, \{f(\cdot, \bar{a})\}_{\bar{a} \in \mathbf{A}^m})$) this result relies on. The Kullback-Liebler divergence, $I(f(\cdot), g(\cdot)) = \int_{\mathbb{R}^d} f(x) \log \frac{f(x)}{g(x)} dx$, is a widely used measure of dissimilarity between two distributions. We denote $I(\bar{a}, \bar{b}) = I(f(\cdot, \bar{a}), f(\cdot, \bar{b}))$ for brevity. We assume that our space of distributions $\{f(\cdot, \bar{a})\}_{\bar{a} \in \mathbf{A}^m}$ satisfies the condition $0 < I(\bar{a}, \bar{b}) < \infty$ for any different $\bar{a}, \bar{b} \in \mathbf{A}^m$. Following [3], we consider only *uniformly good strategy*, i.e., the ones with the cumulative regret of order $o(T^\alpha)$ for any $\alpha > 0$ and any $A \in \mathbf{A}^N$. Assume, WLOG, that each of arms $1, \dots, m, \dots, n$ is included in at least one optimal list (one with the highest reward expectation $\max_{\pi \in \Pi} \mathbb{E}R(\bar{a}_\pi)$) and each of arms $n+1, \dots, N$ is not. We call arms from these two groups *relevant* and *irrelevant* respectively. We denote $\Pi_j := \{\pi \in \Pi : j \in \{\pi(k)\}_{k \in S}\}$, $\bar{a} = (\bar{a}(1), \dots, \bar{a}(m))$ and use $\bar{a}^{\{k \leftarrow a\}}$ for the list of parameter values \bar{a} with a substituted into the position k .

Our first assumption is similar to (but weaker than) the combination of Equations 2.2 and 2.4 from [3].

Assumption 1. *Denseness condition: for any list $\bar{a}_0 \in \mathbf{A}^m$, slot k , finite set of lists $\bar{A} \subset \mathbf{A}^m$, and $\rho > 0$, there exists $a'_0 \in \mathbf{A}$ s.t. (i) $\mathbb{E}R(\bar{a}_0) < \mathbb{E}R(\bar{a}_0^{\{k \leftarrow a'_0\}})$, (ii) for any list $\bar{a} \in \bar{A}$ and slot k' s.t. $\mathbb{E}R(\bar{a}_0^{\{k \leftarrow \bar{a}(k')\}}) \neq \mathbb{E}R(\bar{a}_0)$, we have $I(\bar{a}, \bar{a}^{\{k' \leftarrow a'_0\}}) \leq (1 + \rho)I(\bar{a}, \bar{a}^{\{k' \leftarrow \bar{a}_0(k)\}})$.*

Assumption 1 states that we can improve performance of any list by substituting such a value into an arbitrary position, which is arbitrarily “close” to the replaced value in terms of the reward dis-

tributions if a set of lists. This assumption holds, e.g., if $\mathbf{A} = \mathbb{R}$, function $I(\bar{a}, \bar{b}): \mathbf{A}^{2m} \rightarrow \mathbb{R}$ is continuous and $\mathbb{E}R(\bar{a})$ is strictly monotone with respect to any $\bar{a}(k)$, $k \in S$. Denote by $N_T(\pi)$ the number of times list π is used up to step T . The following lemma provides a lower bound for the regret in an implicit form and helps to obtain an explicit lower bound stated by Theorem 1 under an additional assumption.

Lemma 1. *Under Assumption 1, for any uniformly good strategy and any $A \in \mathbf{A}^N$, for any relevant arm $i \leq n$ and any irrelevant arm j , the set of numbers $\{N_T(\pi)\}_{\pi \in \Pi_j}$ satisfies the following inequality:*

$$\liminf_{T \rightarrow \infty} \sum_{\pi \in \Pi_j} \frac{\mathbb{E}N_T(\pi)}{\log T} I(\bar{a}_\pi, \bar{a}_\pi^{\{\pi^{-1}(j) \leftarrow a_i\}}) \geq 1 \quad (1)$$

Consequently, there exists $x: \mathbb{N} \times \Pi \rightarrow \mathbb{R}_+$ such that, for all $i \leq n, j > n$, we have $\liminf_{T \rightarrow \infty} \sum_{\pi \in \Pi_j} x(T, \pi) I(\bar{a}_\pi, \bar{a}_\pi^{\{\pi^{-1}(j) \leftarrow a_i\}}) \geq 1$ and the cumulative regret over T steps satisfies

$$\liminf_{T \rightarrow \infty} \frac{\text{Reg}_T}{\log T} \geq \liminf_{T \rightarrow \infty} \sum_{\pi \in \Pi} x(T, \pi) \text{Reg}(\pi), \quad (2)$$

where $\text{Reg}(\pi) = \max_{\pi' \in \Pi} \mathbb{E}R(\bar{a}_{\pi'}) - \mathbb{E}R(\bar{a}_\pi)$.

This result generalizes Theorem 3.1 from [3] to two issues: (i) a contribution of each arm to the list reward $R(\bar{a})$ may be not observed; (ii) the reward of the list depends on the order of the arms. Note that Lemma 1 does not use any assumption on the relations between the regret distributions $f(\cdot, \bar{a})$ of different lists \bar{a} , e.g., overlapping in their values. Intuitively, the less relations encoded in the space $(\mathbf{A}, \{f(\cdot, \bar{a})\}_{\bar{a} \in \mathbf{A}^m})$ are, the higher the actual regret of the optimal strategy is (we informally call such relations *correlation*). In fact, the bound from Lemma 1 will be tight only in the case of “full information” (see, e.g., Theorem 3). One can give a formal definition for the opposite case of no information (omitted due to lack of space), when the observed rewards of one list of arms tell nothing about the reward distributions of the others. In this case, our problem setting reduces to the standard stochastic MAB problem with single plays by considering each list as a separate arm. Within it, the tight lower bound for the regret is provided in [12, Theorem 1]. Hence, we return to the setting with Assumption 1.

An explicit bound on the regret can be found as the infimum of the right-hand side of Equation 2 over possible functions $x(T, \pi)$. We claim, omitting a rather standard proof, that there exists $x(T, \pi)$ which provides the minimum and have a finite limit $y(\pi) = \lim_{T \rightarrow \infty} x(T, \pi)$ for each $\pi \in \Pi$. To find the optimal values of $y(\pi)$, we consider the Karush-Kuhn-Tucker conditions for the minimization of the right-hand side of Equation 2 under the constraints defined by Equation 1 for all $i \leq n$ and $j > n$ with $\liminf_{T \rightarrow \infty} \mathbb{E} \frac{N_T(\pi)}{\log T}$ replaced by $y(\pi)$:

$$\begin{cases} \sum_{\pi \in \Pi_j} y_\pi I(\bar{a}_\pi, \bar{a}_\pi^{\{\pi^{-1}(j) \leftarrow a_i\}}) = 1 & \text{or } \lambda_{i,j} = 0 \text{ for any } j > n, i \leq n \\ \sum_{k \in S} \lambda_{i,\pi(k)} I(\bar{a}_\pi, \bar{a}_\pi^{\{k \leftarrow a_i\}}) = \text{Reg}(\pi) & \text{or } y_\pi = 0 \text{ for any } \pi \in \Pi \end{cases} \quad (3)$$

Thus, the optimal values of y_π could be found by comparing solutions of all the linear systems over different arms i, j and lists π satisfying $\lambda_{i,j} = 0$ and $y_\pi = 0$ respectively.

However, the minimum can be found more efficiently under the following assumption about decomposition of a list reward into the sum of the arms' rewards, which is almost always accepted in the literature [3, 17, 9, 16, 18], because it is satisfied by different measures of profit for many applications.

Assumption 2. *Decomposition condition: (i) $R(\bar{a}) = \sum_{k \in S} F(k, \bar{a}(k))$, where $\{F(i, \bar{a}(k))\}_{k \in S}$ are independent, (ii) vector $F(\bar{a})$ includes $F(1, \bar{a}(1)), \dots, F(m, \bar{a}(m))$ as its components.*

For example, most online measures of the web search quality cumulate some relevance gains over documents, e.g., clicks or their dwell times. Observability of the values $F(1, \bar{a}(1)), \dots, F(m, \bar{a}(m))$ (condition (ii)) is crucial for our analysis, since it allows to aggregate information about the plays of an arm in a slot regardless the arms chosen for other slots. We denote the distribution density of $F(k, a)$ by $f(\cdot, k, a)$, introduce $I_k(a, b) := I(f(\cdot, k, a), f(\cdot, k, b))$ and $\text{Reg}(k, j) := \min_{\pi \in \Pi: \pi(k)=j} \text{Reg}(\pi)$, and use A_k^* for the set of arms which are placed in slot k in at least one optimal list. Assumption 2 allows us to present the lower bound for the regret in the following simple form.

Theorem 1. Under Assumptions 1 and 2, for any uniformly good strategy and any $A \in \mathbf{A}^N$,

$$\liminf_{T \rightarrow \infty} \frac{\mathbf{Reg}_T}{\log T} \geq \sum_{j > n} \max_{i \leq n} \min_{k \in S} \frac{\text{Reg}(k, j)}{I_k(a_j, a_i)} \quad (4)$$

This result is very intuitive: the maximization means that we should distinguish an irrelevant arm j from any relevant arm i , the minimization reflects the hope that we are able to make exploratory observations mostly in optimal slots, and the optimized component is standard. Note that Theorem 1 improves only the representation of the lower bound given by Lemma 1 but not the bound itself, what is impossible under Assumptions 1 and 2 (as we prove in Section 4).

On the other hand, adding requirement of the uncorrelation between reward distributions of an arm in different slots allows to obtain a higher lower bound in Theorem 2. The uncorrelation combined with Assumption 1 under Assumption 2 is formalized in the following assumption.

Assumption 3. *Uncorrelation-over-positions denseness condition: for any values $a, a_0 \in \mathbf{A}$, slot k and $\rho > 0$, there exists $a'_0 \in \mathbf{A}$ s.t. (i) $\mathbb{E}F(k, a_0) < \mathbb{E}F(k, a'_0)$, (ii) $I_k(a, a'_0) < (1 + \rho)I_k(a, a_0)$, (iii) for any slot $k' \neq k$, we have $f(\cdot, k', a) = f(\cdot, k', a'_0)$.*

This assumption holds, e.g., if $\mathbf{A} = \mathbb{R}^m$, $F(k, a) = G(a^k)$ for $a = (a^1, \dots, a^m)$, where a space of distributions $\{G(\theta)\}_{\theta \in \mathbb{R}}$ is characterized by a strictly monotone (in θ) expectation function and a continuous function $I(\theta_1, \theta_2)$.

Theorem 2. Under Assumptions 2 and 3, for any uniformly good strategy and any $A \in \mathbf{A}^N$, the number of plays $N_T(k, j)$ of any irrelevant arm $j > n$ in any slot k during the first T steps satisfies the following inequality for any arm $i \in A_k^*$:

$$\liminf_{T \rightarrow \infty} \frac{\mathbb{E}N_T(k, a_j)}{\log T} \geq \frac{1}{I_k(a_j, a_i)}, \quad (5)$$

Then the cumulative regret satisfies the following lower bound:

$$\liminf_{T \rightarrow \infty} \frac{\mathbf{Reg}_T}{\log T} \geq \sum_{j > n, k \in S} \max_{i \in A_k^*} \frac{\text{Reg}(k, a_j)}{I_k(a_j, a_i)} \quad (6)$$

Naturally, in order to distinguish the arm j from the arm i in the slot k , we need to play it in this slot the same number of times as in the standard SMAB problem with one play and the optimal arm i . Though reward distributions of an object in different slots seem to be dependent in practice, it may be of use for constructing a strategy to treat them as independent if the dependence is difficult to be inferred. As an example, one can consider a project with various tasks requiring different competencies and to be assigned to different workers from a big set of candidates, e.g., a football match, where a manager chooses players for different positions. We also note that max in Equation 6 disappears if there is the only optimal list. Now we prove our claims.

Proof of Lemma 1. At the first step of our proof, we use the change of measure technique, like [3] does, and prove that, for any irrelevant arm $j > n$, the vector of numbers $\{N_T(\pi)/\log T\}_{\pi \in \Pi_j}$, with high probability, lays outside of some $|\Pi_j|$ -dimensional cuboids of the form $\{\{x(\pi)\}_{\pi \in \Pi_j} : 0 \leq x_\pi < c(\pi)\}$. At the second step, which is completely novel and crucial for the new issues, we aggregate these estimates to show that this vector is outside of a sequence of simplexes what in the limit provides Equation 1.

Step 1. Consider any optimal arm list π_0 , any arm $i \in \pi_0(S)$, and any irrelevant arm $j > n$. According to Assumption 1 applied to the list π_0 , the slot $\pi_0^{-1}(i)$ and the set of lists Π_j , for a fixed $\rho > 0$, we can choose a value $a^* \in \mathbf{A}$ such that

$$(i) \mathbb{E}R(\bar{a}_{\pi_0}^{\{\pi_0^{-1}(i) \leftarrow a^*\}}) > \mathbb{E}R(\bar{a}_{\pi_0}), (ii) (1 + \rho)I(\bar{a}_\pi, \bar{a}_\pi^{\{\pi^{-1}(j) \leftarrow a_i\}}) > I(\bar{a}_\pi, \bar{a}_\pi^*) \quad \forall \pi \in \Pi_j, \quad (7)$$

where we denote $\bar{a}_\pi^* := \bar{a}_\pi^{\{\pi^{-1}(j) \leftarrow a^*\}}$ for $\pi \in \Pi_j$. We use the “alternative” parameter values $A^* = \{a_1, \dots, a_{j-1}, a^*, a_{j+1}, \dots, a_N\}$ to prove the following statement.

Lemma 2. Consider any $\bar{c} = \{c(\pi)\}_{\pi \in \Pi_j}$ satisfying $\sum_{\pi \in \Pi_j} c(\pi)I(\bar{a}_\pi, \bar{a}_\pi^*) = \delta < \frac{1}{1+\rho}$. We have

$$\lim_{T \rightarrow \infty} P_A \left(\prod_{\pi \in \Pi_j} \{N_T(\pi)/\log T < c(\pi)\} \right) = 0. \quad (8)$$

The proof is based on the log odds ratio of the likelihood of the rewards $R_1(\bar{a}_\pi), \dots, R_t(\bar{a}_\pi)$ (observed at t plays of a list π) under the parameter values \bar{a}_1 and \bar{a}_2 : $L_{t,\pi}(\bar{a}_1, \bar{a}_2) = \sum_{\tau=1}^t \log \frac{f(R_\tau(\bar{a}_\pi), \bar{a}_1)}{f(R_\tau(\bar{a}_\pi), \bar{a}_2)}$. By the strong law of large numbers, we have

$$I(\bar{a}_\pi, \bar{a}_\pi^*) = \mathbb{E}_A \left(\log \frac{f(R(\bar{a}_\pi), \bar{a}_\pi)}{f(R(\bar{a}_\pi), \bar{a}_\pi^*)} \right) = \lim_{t \rightarrow \infty} \frac{L_{t,\pi}(\bar{a}_\pi, \bar{a}_\pi^*)}{t} = \lim_{t \rightarrow \infty} \frac{\max_{\tau \leq t} L_{\tau,\pi}(\bar{a}_\pi, \bar{a}_\pi^*)}{t} \quad P_A\text{-a.s.}$$

Consequently, for $\pi \in \Pi_j$ and events $B_{\tau,T}^{\pi,c} := \{L_{\tau,\pi}(\bar{a}_\pi, \bar{a}_\pi^*) \leq (1+\rho)I(\bar{a}_\pi, \bar{a}_\pi^*)c \log T\}$, we have

$$\lim_{T \rightarrow \infty} P_A \left(\prod_{\tau < c \log T} B_{\tau,T}^{\pi,c} \right) = 1. \quad (9)$$

Using Equation 9, we obtain Lemma 2 in the following way:

$$\begin{aligned} \lim_{T \rightarrow \infty} P_A \left(\prod_{\pi \in \Pi_j} \left\{ \frac{N_T(\pi)}{\log T} < c(\pi) \right\} \right) &\leq \lim_{T \rightarrow \infty} \left[P_A \left(\prod_{\pi \in \Pi_j} \left(\left\{ \frac{N_T(\pi)}{\log T} < c(\pi) \right\} \prod_{\tau < c(\pi) \log T} B_{\tau,T}^{\pi,c(\pi)} \right) \right) \right] + \\ &+ 1 - P \left(\prod_{\pi \in \Pi_j} \prod_{\tau < c(\pi) \log T} B_{\tau,T}^{\pi,c(\pi)} \right) \leq \lim_{T \rightarrow \infty} P_A \left(\prod_{\pi \in \Pi_j} \left\{ \frac{N_T(\pi)}{\log T} < c(\pi) \right\} B_{N_T(\pi),T}^{\pi,c(\pi)} \right) = 0. \end{aligned} \quad (10)$$

To prove the last equality, we introduce, for any $\bar{\tau} = \{\tau(\pi)\}_{\pi \in \Pi_j}$ s.t. $\tau(\pi) < c(\pi)$, event $S_j(T, \bar{c}, \bar{\tau}) := \prod_{\pi \in \Pi_j} (\{N_T(\pi) = \tau(\pi)\} B_{N_T(\pi),T}^{\pi,c(\pi)})$ and find:

$$\begin{aligned} P_{A^*}(S_j(T, \bar{c}, \bar{\tau})) &= \int \mathbf{1}\{S_j(T, \bar{c}, \bar{\tau})\} dP_{A^*} = \int \mathbf{1}\{S_j(T, \bar{c}, \bar{\tau})\} e^{-\sum_{\pi \in \Pi_j} L_{\tau(\pi),\pi}(\bar{a}_\pi, \bar{a}_\pi^*)} dP_A \geq \\ &\geq T^{-(1+\rho) \sum_{\pi \in \Pi_j} c(\pi) I(\bar{a}_\pi, \bar{a}_\pi^*)} P_A(S_j(T, \bar{c}, \bar{\tau})) \geq T^{-\delta(1+\rho)} P_A(S_j(T, \bar{c}, \bar{\tau})), \end{aligned} \quad (11)$$

where the second equality uses the change of measure, which concerns only arm j , and the third inequality is based on the definition of $B_{N_T(\pi),T}^{\pi,c(\pi)}$. Since $\prod_{\pi \in \Pi_j} \{N_T(\pi) < c(\pi) \log T\} B_{N_T(\pi),T}^{\pi,c(\pi)} = \bigcup_{\bar{\tau}: \tau(\pi) < c(\pi) \log T} \prod_{\pi \in \Pi_j} S_j(T, \bar{c}, \bar{\tau})$, where united sets are disjoint, we obtain from Equation 11

$$P_A \left(\prod_{\pi \in \Pi_j} \left\{ \frac{N_T(\pi)}{\log T} < c(\pi) \right\} B_{N_T(\pi),T}^{\pi,c(\pi)} \right) \leq T^{\delta(1+\rho)} P_{A^*} \left(\prod_{\pi \in \Pi_j} \left\{ \frac{N_T(\pi)}{\log T} < c(\pi) \right\} B_{N_T(\pi),T}^{\pi,c(\pi)} \right) \quad (12)$$

Equation 7 (i) implies that, under A^* , any optimal list π_{opt} belongs to Π_j . Since the strategy is uniformly good, we also have $P_{A^*} \left\{ \frac{N_T(\pi_{opt})}{\log T} < c \right\} \leq \frac{\mathbb{E}_{A^*}(T - N_T(\pi_{opt}))}{T - c \log T} = \frac{o(T^\alpha)}{T - c \log T} = o(T^{\alpha-1})$ for any $c > 0$. Therefore, Equation 12 implies that its left-hand side is $o(T^{\alpha-1+\delta(1+\rho)}) = o(1)$, if we choose $\alpha \in (0, 1 - \delta(1+\rho))$.

Step 2. Fix any $\epsilon > 0$ and choose $\rho > 0$ such that $\frac{1}{(1+2\rho)(1+\rho)} > 1 - \epsilon$. We obtain Equation 1 from

$$\mathbb{E} \frac{\sum_{\pi \in \Pi_j} N_T(\pi) I(\bar{a}_\pi, \bar{a}_\pi^{\{\pi^{-1}(j) \leftarrow a_i\}})}{\log T} \geq P_A \left(\frac{\sum_{\pi \in \Pi_j} N_T(\pi) I(\bar{a}_\pi, \bar{a}_\pi^{\{\pi^{-1}(j) \leftarrow a_i\}})}{\log T} \geq 1 - \epsilon \right) (1 - \epsilon) \xrightarrow{T \rightarrow \infty} 1 - \epsilon.$$

This convergence is equivalent to

$$\lim_{T \rightarrow \infty} P_A \left(\{N_T(\pi) / \log T\}_{\pi \in \Pi_j} \in S_{1-\epsilon} \right) = 0 \quad (13)$$

for the simplex $S_{1-\epsilon} = \left\{ \{x(\pi)\}_{\pi \in \Pi_j} : \sum_{\pi \in \Pi_j} x(\pi) I(\bar{a}_\pi, \bar{a}_\pi^{\{\pi^{-1}(j) \leftarrow a_i\}}) < 1 - \epsilon, x(\pi) \geq 0 \right\}$. Note that it can be covered by a finite union of cuboids $C_{\{c(\pi)\}_{\pi \in \Pi_j}} = \left\{ \{x(\pi)\}_{\pi \in \Pi_j} : 0 \leq x(\pi) < c(\pi) \right\}$ contained in $S_{\frac{1}{(1+2\rho)(1+\rho)}}$, i.e., satisfying $\sum_{\pi \in \Pi_j} c(\pi) I(\bar{a}_\pi, \bar{a}_\pi^{\{\pi^{-1}(j) \leftarrow a_i\}}) < \frac{1}{(1+2\rho)(1+\rho)}$. Due to Equation 7, the latter condition

implies $\sum_{\pi \in \Pi_j} c(\pi) I(\bar{a}_\pi, \bar{a}_\pi^*) < \frac{1}{1+2\rho}$. Then, from Equation 8, for each of these cuboids, $\lim_{T \rightarrow \infty} P_A \left(\{N_T(\pi)/\log T\}_{\pi \in \Pi_j} \in C_{\{c(\pi)\}_{\pi \in \Pi_j}} \right) = 0$ what implies Equation 13.

Finally, Equation 1.1 from [12] yields $\liminf_{T \rightarrow \infty} \frac{\text{Reg}_T}{\log T} = \liminf_{T \rightarrow \infty} \sum_{\pi \in \Pi} \mathbb{E} N_T(\pi) \text{Reg}(\pi) / \log T$, and we obtain Equation 2 by minimizing this expression over $\{N_T(\pi)\}_{T \in \mathbb{N}, \pi \in \Pi}$ satisfying Equation 1 for all $i \leq n, j > n$. Existence of an optimal function $x(T, \pi)$ is discussed above, before Equation 3. \square

Proof of Theorem 1. Under Assumption 2, we have $f((x^1, \dots, x^m), \bar{a}) = f(x^1, \bar{a}(1)) \dots f(x^m, \bar{a}(m))$, and, thus, for any list $\bar{a} \in \mathbf{A}^m$, slot k and value $a^* \in \mathbf{A}$,

$$I(\bar{a}, \bar{a}^{\{k \leftarrow a^*\}}) = \int_{\mathbb{R}^m} f(x^1, \bar{a}(1)) \dots f(x^m, \bar{a}(m)) \log \frac{f(x^k, \bar{a}(k))}{f(x^k, a^*)} dx^1 \dots dx^m = I_k(\bar{a}(k), a^*) \quad (14)$$

Then, we can rewrite Equation 1 as follows:

$$\liminf_{T \rightarrow \infty} \mathbb{E} \sum_{k \in S} N_T(k, j) I_k(a_j, a_i) / \log T \geq 1 \quad (15)$$

For further simplification of these restrictions, we utilize the following combinatorial lemma which claims that, in order to minimize regret under the fixed values of $\{N_T(k, j)\}_{j > n, k \in S}$, a strategy should not observe several arms $j > n$ at one step.

Lemma 3. *There are m slots and m objects with some reward $r(k, j)$ corresponding to an object j put in a slot k . Let consider $t \leq m$ steps and a subset of different slots $\{k_1, \dots, k_t\}$. Assume we should close each of these slots at exactly one step. After it, at each step, we choose such a combination of different objects to put them in open slots (only one object in one slot) that maximizes the cumulative reward on this step. Then, one of the ways to reach the maximum cumulative reward over all the steps is to close one slot per step.*

Proof sketch. The idea of the proof is that, when closing just one slot at each step, we can repeat any combination $\{N_T(k, j)\}_{k=1, \dots, m, j=1, \dots, m}$ which can be reached by any other strategy. We drop the accurate proof due to its technical nature. \square

Then, while considering only rational strategies from Lemma 3, each play of the arm j in the slot k corresponds to a step with regret not less than $\text{Reg}(k, j)$, what leads to the following estimate:

$$\begin{aligned} \liminf_{T \rightarrow \infty} \text{Reg}_T / \log T &\geq \liminf_{T \rightarrow \infty} \sum_{j > n} \sum_{k \in S} N_T(k, j) \text{Reg}(k, j) / \log T = \\ (\forall \{i_j\}_{j > n}, i_j \leq n) &= \liminf_{T \rightarrow \infty} \sum_{j > n} \sum_{k \in S} \frac{N_T(k, j) I_k(a_j, a_{i_j})}{\log T} \frac{\text{Reg}(k, j)}{I_k(a_j, a_{i_j})} \geq \\ &\geq \liminf_{T \rightarrow \infty} \sum_{j > n} \left[\min_{k \in S} \frac{\text{Reg}(k, j)}{I_k(a_j, a_{i_j})} \cdot \sum_{k \in S} \frac{N_T(k, j) I_k(a_j, a_{i_j})}{\log T} \right] \geq \sum_{j > n} \min_{k \in S} \frac{\text{Reg}(k, j)}{I_k(a_j, a_{i_j})}, \end{aligned}$$

where the last inequality follows from Equation 15. Taking maximum over all possible sets $\{i_j\}_{j > n}$ yields Equation 4. \square

Proof of Theorem 2. We describe a modification of Step 1 of the proof of Lemma 1 which proves the current theorem. Given an irrelevant arm $j > n$ and an optimal list π_0 with an arm i in a slot k , according to Assumption 3, we can choose such a value $a^* \in \mathbf{A}$ that

$$\mathbb{E} F(k, a_i) < \mathbb{E} F(k, a^*), I_k(a_j, a^*) < (1 + \rho) I_k(a_j, a_i), \forall k' \neq k f(\cdot, k', a_j) = f(\cdot, k', a^*) \quad (16)$$

Then, in the case of the arm parameters $A^* = \{a_1, \dots, a_{j-1}, a^*, a_{j+1}, \dots, a_N\}$, the list π_0 is optimal and, since the probability to observe a fixed reward at some step in some slot differs under measures P_A and P_{A^*} only if the slot is k with arm j in it at this step, we can estimate each value $N_T(k, j), k \in S$, separately. Indeed, by choosing $c(\pi) \leq \frac{1}{(1+2\rho)I_k(a_j, a^*)}$ for each list π with $\pi(k) = j$ and putting $c(\pi) = +\infty$ for other lists, we obtain estimates analogous to Equations 11–8 resulting in $\lim_{T \rightarrow \infty} P_A \left(N_T(k, j) < \frac{\log T}{(1+2\rho)I_k(a_j, a^*)} \right) = 0$ at the end of Step 1 of the proof of Lemma 1. Applying the first inequality from Equation 16 and letting $\rho \rightarrow 0$ yields Equation 5. Maximizing its right-hand side over $i \in A_k^*$ leads to Equation 6. \square

4 Asymptotically Optimal Algorithm

In this section, we construct algorithms with asymptotically optimal regret reaching lower bounds from Lemma 1 and Theorems 1 and 2. In our construction, we rely on the algorithm proposed in [3] under Assumptions 1 and 2 with additional constraint $f(\cdot, i, a) = f(\cdot, a)$ and modify it to handle the case of non-equivalent plays. First, we supplement the setting of Theorem 1 with the following assumption under which we are able to present an optimal algorithm.

Assumption 4. *Factorization condition: the arm reward has a form $F(k, a) = p(k)r(a)$, where $p(k)$ is a Bernoulli random variable with a parameter dependent on k only, $r(a)$ is a random variable with distribution dependent on a only. Besides, values of $p(k)$, $k \in S$, are components of $F(\bar{a})$.*

This factorization model is often used in different applied problems, e.g., it corresponds to the examination hypothesis [8] underlying different models of user behavior on the web search result page. Under this hypothesis, the variable $p(k)$ indicates whether the user examined the document, and $r(a)$ measures the user satisfaction with it. One of possible ways to observe these values is to consider a click or a click on a lower position as a fact of examination and a satisfied click (e.g., one with a long enough dwell time or the last click in the session [7, 5]) as a fact of user satisfaction. More general but similar factorization assumption was also considered in [17].

We denote $\mathbb{E}p(k) = p_k$, $\mathbb{E}r(a_j) = \mu_j$ and assume, WLOG, that $\mu_{a_1} \geq \dots \geq \mu_{a_l} > \mu_{a_{l+1}} = \dots = \mu_{a_m} = \dots = \mu_{a_n} > \mu_{a_{n+1}} \geq \dots \geq \mu_{a_N}$ and the slots $1, \dots, m$ are ordered by decreasing p_k . Below we assume that the agent knows this order. Otherwise, it could sort the slots by current empirical estimates of p_k . Errors of these estimates will not influence on the asymptotic behavior of the regret, due to the exponential convergence rate of the mean estimate provided by the Chernoff-Hoeffding bound: for iid random variables x_1, \dots, x_n with values in $[0, 1]$ and for any $\epsilon > 0$, $P((x_1 + \dots + x_n)/n - \mathbb{E}x_i < -\epsilon) \leq e^{-2K\epsilon^2}$.

Due to Assumption 4, the value of $r(a)$ is observed only if $p(k) = 1$ for the corresponding slot. Further, when it is observed, its distribution does not depend on the slot. Then, we define arm-dedicated statistics $\mu_{j,t}$ and $U_{j,t}$ from [3] which, in our case, are based not on all the plays of the arm j but only on all the observations of $r(a_j)$. First one $\mu_{j,t}$ estimates expectation μ_j of $r(a_j)$: $\mu_{j,t} = \sum_{i=1}^{N_t^*(j)} r_i(a_j) / N_t^*(j)$, where $N_t^*(j)$ is the number of observations of $r(a_j)$ during the first t steps and $r_i(a_j)$ is the i -th observed value. The second statistics $U_{j,t} = g_{t, N_t^*(j)}(r_1(a_j), \dots, r_{N_t^*(j)}(a_j))$ (see definition of $g_{t,s}(Y_1, \dots, Y_s)$ in Section IV of [3]; we define Y_i as an observation of $r(a)$ for some a) is a kind of an upper confidence bound used in different MAB algorithms, e.g., UCB-1 [4] Bayesian-UCB [10] and is constructed to satisfy the asymptotic properties proved in Theorem 4.2 in [3] under the following assumption.

Assumption 5. (i) *The space of reward distributions can be parametrized by $\theta \in \mathbb{R}$, i.e., $f(\cdot, k, a) = f(\cdot, \theta)$, in such a way that $\log f(x, \theta)$ is concave in θ for each x . (ii) $\int x^2 f(x, \theta) dx < \infty$.*

Based on the statistics $\mu_{j,t}$ and $U_{j,t}$, we describe an asymptotically optimal algorithm under Assumptions 1, 2, 4 and 5 in Algorithm 1. Given values of the statistics, it chooses m arms to be observed as the algorithm from [3] does it and ranks them by decreasing $\mu_{j,t}$. Theorem 3 states its optimality.

Theorem 3. *Algorithm 1 is asymptotically optimal under Assumptions 1, 2, 4 and 5, i.e., the asymptotics of its regret coincides with the lower bound from Theorem 1: $\liminf_{t \rightarrow \infty} \frac{\text{Reg}_t}{\log t} =$*

$$\sum_{j>n} \frac{\mu_{a_m} - \mu_{a_j}}{p_m I(a_j, a_m)}.$$

Proof. The following estimates show that, under Assumption 4, the latter asymptotics corresponds to the lower bound: $I_k(a_j, a_i) = p_k I(a_j, a_i) \geq p_k I(a_j, a_m)$,

$$\begin{aligned} \frac{\text{Reg}(k, a_j)}{p_k} &= \frac{(\mu_{a_k} - \mu_{a_j})(p_k - p_{k+1}) + \dots + (\mu_{a_{m-1}} - \mu_{a_j})(p_{m-1} - p_m) + (\mu_{a_m} - \mu_{a_j})p_m}{p_k} \geq \\ &= \frac{(\mu_{a_m} - \mu_{a_j})((p_k - p_{k+1}) + \dots + (p_{m-1} - p_m) + p_m)}{p_k} = \mu_{a_m} - \mu_{a_j}. \end{aligned}$$

Further proof differs from that of Theorem 5.1 from [3] by the two issues: (i) an optimal arm list with a suboptimal order of arms provides the zero regret in the original case and a non-zero one in our case; (ii) in our case, a play of an arm j does not necessarily provide an observation of $r(a_j)$. Our proof consists of the following steps corresponding to steps from [3].

Algorithm 1: Asymptotically optimal bandit algorithm under Assumptions 1, 2, 4 and 5

Data: m , space $(\mathbf{A}, \{f(\cdot, \bar{a})\}_{\bar{a} \in \mathbf{A}^m})$, arm parameters A , slots S ;

- 1 Make m observations of r_j (with $p(k) = 1$) for each arm j (in any slots); $t_0 \leftarrow \#$ of steps for it;
- 2 Choose $\delta \in (0, p_m/(2N^2))$;
- 3 **for** $t = t_0$ **to** T **do**
- 4 $j^* \leftarrow t \% N$; // choose an arm uniformly over steps; $x \% y$ is a remainder of division of x by y ;
- 5 $G \leftarrow \emptyset$; **for** $j = 1$ **to** N **do**
- 6 **if** $N_t^*(j) > \delta t$ **then** $G \leftarrow G \cup \{j\}$;
- 7 **if** $|G| < m$ **then** Add different $(m - |G|)$ arms to G randomly;
- 8
- 9 $\pi'_t \leftarrow$ list of top- m arms from G by decreasing $\mu_{j,t}$;
- 10 **if** $j^* \in \pi'_t(S)$ **then**
- 11 Show $\pi_t := \pi'_t$;
- 12 **else**
- 13 **if** $U_{j^*,t} < \mu_{\pi'_t(m),t}$ **then** Show $\pi_t := \pi'_t$; **else** Show $\pi_t := (\pi'_t(1), \dots, \pi'_t(m-1), j^*)$;
- 14 Observe user feedback $F(\bar{a}_{\pi_t})$;

Result: Arm list at each step: $\{\pi_t\}_{t=1, \dots, T}$

- **Step A.** For each relevant arm $i \leq l$, we have $\mathbb{E}N_T(i, i) = T - o(\log T)$.
- **Step B.** $\mathbb{E}B_T = T - o(\log T)$, where $B_T = \#\{t \leq T \mid \pi'_t \text{ consists only of arms } j \leq n\}$, where π'_t is defined in line 8 of Algorithm 1.
- **Step C.** For any $j > n$ and $\rho > 0$ there exists $\epsilon > 0$ such that $\mathbb{E}S_T(j) \leq \frac{1+\rho+o(1)}{I_m(a_j, a_m)} \log T$, where

$$S_T(j) = \#\{t \leq T \mid \pi'_t(i) = i \ \forall i \leq l, \ |\mu_{i,t} - \mu_i| < \epsilon \ \forall i \leq n, \ \pi'_t \text{ consists only of arms } i \leq n \text{ and } r(a_j) \text{ is observed at step } t\}.$$

Now we explain how Steps A, B and C are combined to yield Theorem 3. Let consider the following particular case of the Chernoff-Hoeffding bound for independent observations of the indicator of an $r(a_j)$ observation given an arm j is played at a particular step:

$$P(N_t^*(j) < N_t(j)p_m/2) \leq e^{-N_t(j)p_m^2/2}, \quad (17)$$

where $N_t(j)$ is the number of plays of the arm j during the first t steps. Along with the condition $\delta < p_m/(2N^2)$, this estimate implies that both the expected number of steps of Algorithm 1 with active line 7 and the expected number of steps in line 1 are finite. Combined with Steps A and B, this provides that the cumulative regret of Algorithm 1 is of order $o(\log T)$, except for steps counted by $S_T(j), j > n$ at Step C. The regret at these steps is at most

$$\sum_{j>n} \frac{(\mu_{a_m} - \mu_{a_j})(1 + \rho + o(1))}{I_m(a_j, a_m)} \log T = \sum_{j>n} \frac{(\mu_{a_m} - \mu_{a_j})(1 + \rho)}{p_m I(a_j, a_m)} \log T + o(\log T)$$

Letting $\rho \rightarrow 0$ concludes the proof. \square

Proof of Step A. Choose $c > (N+1)(1 - 2N^2\delta/p_m)^{-1}$ to provide $[(c^r - c^{r-1})/N] > 2N\delta c^r/p_m$ for $r \in \mathbb{N}$ and choose $\epsilon < \min\{(\mu_{a_i} - \mu_{a_j})/2, i < j \leq l; (\mu_{a_l} - \mu_{a_m})/2; (\mu_{a_m} - \mu_{a_{n+1}})/2\}$. Lemmas 5.1 and 5.2 and their proofs could be transferred from [3] to our case without any changes. Now we change Lemma 5.3 from [3] by the following extended analysis. By Lemma 5.2, on $A_r B_r$, any arm $i \leq l$ satisfies $N_t(i) \geq [(c^r - c^{r-1})/N] > 2\delta t/p_m$ for any step $t \in [c^r, c^{r+1}]$ and $r \geq r^*$ for some r^* . Then, we estimate $N_t^*(i)$ by Equation 17: $P(N_t^*(i) < \delta t \mid A_r B_r) \leq e^{-\delta t p_m}$. In combination with Lemma 5.1, it implies $P(C_r) = 1 - o(c^{-r})$ for $C_r = \prod_{i \leq l} \{N_t^*(i) \geq \delta t\} A_r B_r$. Further, at step t , on C_r , each arm $i \leq l$ is included in π'_t in line 6 and, moreover, $\pi'_t(i) = i$, i.e., it is played at its optimal position, since on A_r Algorithm 1 sorts π'_t perfectly in line 8. Finally, we can estimate $\mathbb{E}N_t(i, i) \geq \sum_{r=r_0}^{\lfloor \log t \rfloor} \sum_{c^r \leq t < \min\{c^{r+1}, t\}} P(C_r) = \sum_{r=r_0}^{\lfloor \log t \rfloor} (\min\{c^{r+1}, t\} - c^r - o(1)) = t - o(\ln t)$. \square

Proof of Step B. Again, we transfer from [3] without changes Lemmas 5.1.B and 5.2.B and the claim proved just after the proof of Lemma 5.2.B. Then, we apply the same trick as at Step A. \square

Proof of Step C. Note that each observation of $r(a_j)$ counted by $S_t(j)$ occurs in the slot m . Then, we transfer the proof of Step C from [3] to our case by changing the notion of a play of the arm j by the notion of an observation of $r(a_j)$. \square

Thus, we proved the tightness of the lower bound for the asymptotic behavior of regret provided by Lemma 1 and Theorem 1. Under Assumptions 2, 3 and 5, a construction of an optimal algorithm reaching the lower bound from Theorem 2 could be similar. Specificity is that (i) the agent should maintain statistics $\mu_{k,j,t}$ and $U_{k,j,t}$ for each pair of a slot k and an arm j and (ii) if, at some step, the agent decides to substitute the arm from the special arm-slot pair for the arm j greedily chosen for this slot, it should find a greedy-optimal combination of arms for other slots again since it may now include j .

5 Conclusion

In this paper, we systematically studied the stochastic non-contextual multi-armed bandit problem with non-equivalent multiple plays. We considered some of the most interesting and, at the same time, quite general problem settings which are covered by our formulation and handle many applied problems. For them, we provided lower bounds for asymptotic behavior of the regret and proved tightness of these bounds. We believe that this work could be a basis both for finding theoretically optimal algorithms in more specific cases of our problem settings and for future development of applied algorithms.

References

- [1] N. Ailon, K. Hatano, and E. Takimoto. Bandit online optimization over the permutahedron. In *Algorithmic Learning Theory*, pages 215–229. Springer, 2014.
- [2] A. Anagnostopoulos, A. Z. Broder, E. Gabrilovich, V. Josifovski, and L. Riedel. Just-in-time contextual advertising. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 331–340. ACM, 2007.
- [3] V. Anantharam, P. Varaiya, and J. Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: Iid rewards. *Automatic Control, IEEE Transactions on*, 32(11):968–976, 1987.
- [4] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.
- [5] P. N. Bennett, F. Radlinski, R. W. White, and E. Yilmaz. Inferring and using location metadata to personalize web search. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’11, pages 135–144, New York, NY, USA, 2011. ACM.
- [6] N. Cesa-Bianchi and G. Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.
- [7] K. Collins-Thompson, P. N. Bennett, R. W. White, S. de la Chica, and D. Sontag. Personalizing web search results by reading level. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM ’11, pages 403–412, New York, NY, USA, 2011. ACM.
- [8] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM ’08, pages 87–94, New York, NY, USA, 2008. ACM.
- [9] S. Kale, L. Reyzin, and R. E. Schapire. Non-stochastic bandit slate problems. In *Advances in Neural Information Processing Systems*, pages 1054–1062, 2010.
- [10] E. Kaufmann, O. Cappe, and A. Garivier. On bayesian upper confidence bounds for bandit problems. In N. D. Lawrence and M. A. Girolami, editors, *AISTATS-12*, volume 22, pages 592–600, 2012.
- [11] W. M. Koolen, M. K. Warmuth, and J. Kivinen. Hedging structured concepts. In *COLT*, pages 93–105. Citeseer, 2010.
- [12] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.
- [13] L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. WWW ’10, pages 661–670, New York, NY, USA, 2010. ACM.
- [14] S. Pandey and C. Olston. Handling advertisements of unknown quality in search advertising. In *Advances in Neural Information Processing Systems*, pages 1065–1072, 2006.
- [15] A. Slivkins, F. Radlinski, and S. Gollapudi. Ranked bandits in metric spaces: Learning diverse rankings over large document collections. *J. Mach. Learn. Res.*, 14(1):399–436, Feb. 2013.
- [16] M. Sloan and J. Wang. Iterative expectation for multi period information retrieval. In *WSDM Workshop on Web Search Click Data*, 2013.
- [17] H. P. Vanchinathan, I. Nikolic, F. De Bona, and A. Krause. Explore-exploit in top-n recommender systems via gaussian processes. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 225–232. ACM, 2014.
- [18] A. Vorobev, D. Lefortier, G. Gusev, and P. Serdyukov. Gathering additional feedback on search results by multi-armed bandits with respect to production ranking. In *Proceedings of the 24th international conference on World wide web*, pages 1177–1187. International World Wide Web Conferences Steering Committee, 2015.